IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 1, Ser. 1 (Jan. – Feb. 2025), PP 56-63
www.iosrjournals.org

An Investigation On Enhancing Student Performance Prediction With Machine Learning Models

Somil Jain, Prof. O. P. Karada, Dr. Deepak Kumar Yadav
Research Scholar, Department Of Computer Science And Engineering, Institute Of Engineering And Technology, SAGE University, Indore, Madhya Pradesh, India
Professor, Department Of Computer Science And Engineering, Institute Of Engineering And Technology, SAGE University, Indore, Madhya Pradesh, India
Head Of Department, Department Of Computer Science And Engineering, Institute Of Engineering And Technology, SAGE University, Indore, Madhya Pradesh, India

Abstract
This study explores the application of machine learning (ML) techniques and Educational Data Mining (EDM) methodologies to predict and enhance student performance. EDM integrates data mining models with educational datasets to extract meaningful insights that optimize the teaching-learning process. By focusing on innovative approaches such as Ant Colony Optimization (ACO) and Logistic Regression (LR), this research develops a hybrid ACO-LR model for feature selection and classification. The findings demonstrate significant improvements in prediction accuracy, precision, and recall, highlighting the effectiveness of ML techniques in addressing educational challenges like student dropouts and performance variability. This study underscores the transformative potential of data-driven decision-making in the educational domain and provides actionable insights for improving student outcomes.

Keywords: *Optimization, Machine Learning, Performance, Education, Data Mining*

Date of Submission: 06-01-2025 Date of Acceptance: 16-01-2025

I. Introduction

The educational system is the backbone of India's economy and social progress. The primary goal of an educational system is to endow students with the knowledge and good employability skills. The recent advancements in technology led most of the educational institutions to step towards Educational Data Mining (EDM). EDM is the method of applying the data mining (DM)methodologies on a large number of student details with the intention of extracting meaningful details that helps in enhancing the teaching-learning process. DM approaches gained significant attention in the educational sector. Decision making and prediction, being the important functionalities of EDM, it is rapidly gaining its importance in the educational field. Particularly, DM models offer the educational plan makers with data-based approaches required to support the motto of enhancing the effectiveness and excellence of the teaching- learning process. In this view, the usage of DM models seems to bring significant changes in the global educational system. EDM assists educational institutions in offering solutions for particular problems.

Educational Data Mining (EDM)

In the last decades, EDM has attained massive attention from researchers due to the existence of massive educational details which is accessible from many sources. The main aim of EDM is to make DM models more effectively in order to safeguard the numerous amounts of educational information and to develop a protective atmosphere for the student's learning. In this approach, diverse models have been deployed for DM and its analytics (Baker *et al.,* 2014). Moreover, prediction models were used namely, Classification, Regression, and Latent factor evaluation technologies.

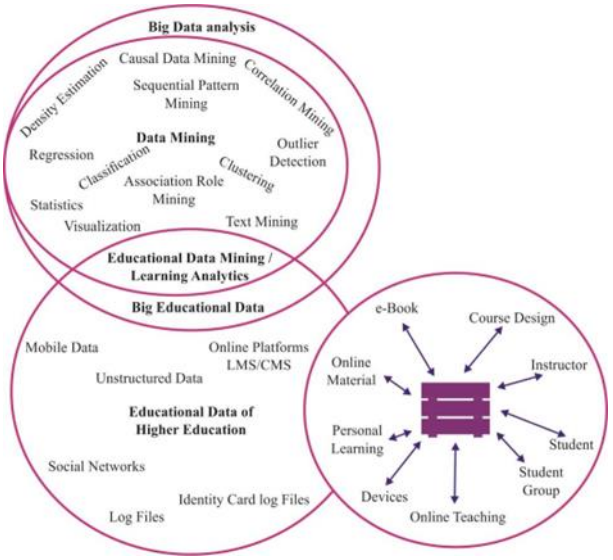

Figure 1. An illustration of DM use in HE

The DM tools are collaborated with academics in enhancing the students learning methodologies by exploring, filtering, and estimating the parameters relevant to student's features or behaviours (Baradwaj *et al.*, 2012) (Figure 1). The major challenges faced by any educational institutions lies in the number of placements it gets and the number of successful graduates it produces.

II. Literature Review

Ray *et al*. (2018) concentrated on the application of DM and data analytics for handling the data produced from the educational industry. The EDM and LA methodologies were defined for managing the big data in commercial as well as alternate communities. Also, it provides an extensive definition in EDM and LA which affects the function of shareholders in the PG level education center. Moreover, a brief description of applying these models, and examining the learning process of students, assessing the performance to provide extensive feedback practically. Eventually, these models affect the administrative principles which are qualified for all stakeholders in an educational institution.

Hegde *et al.* (2018) a model has been developed for predicting the student dropout under the application of NB Classification in R language. Followed by, it examines the root cause for student failure or completion in the first year and analyzes whether the student would be dropped out or not. Massive factors were considered in this study which impacts an immediate dropout as mentioned above. Previous dropout prediction is highly helpful for the firms to maintain the student's academic program.

In (Atallah, *et al.*, 2014) the domains of EDM are to filter the meaningful data from the student's registration form or admission form. The sample applied in this dataset has 5 years period [2005-2011] by offering the analytical tool for reviewing the details for making decision practically under the consideration of grade and GPA of students. In (Agaoglu, 2016) 4 various classifiers have been employed namely DT, SVM, Artificial Neural Networks (ANN), and discriminant analysis (DA). The classification function has been compared with a data set with the student's response for real-time course questionnaire interms of accuracy, precision, recall, and specificity metrics. Even though the above- mentioned classification methods are effective, the C5.0 classifier is assumed to be the optimal one with remarkable accuracy.

Alsuwaiket *et al.* (2020) various data preparation process has been deployed with massive student records for preparing the students marks according to the assessment modules. Here, data is computed under various phases for extracting the categorical factor where student's marks have been refined in the data preparation state. Consequently, the final marks of students are not isolated from the enrolled modules. Followed by, the examination of EDM data pre-processing phases has been performed. Typically, it is finalized that educational information should not be developed similarly as alternate data types because of the variations like data sources, applications, and errors involved in them. Hence, coursework estimation ratio has been employed for considering various module assessment approaches at the time of preparing student transcription data. The impact of coursework assessment ratio (CAR) on detection by applying RF classifier has also been presented.

Gibson *et al.* (2017) DM has gained massive attention from the developers, and model- relied methods, ML, and data science as a novel toolkit for the upcoming generation and addition of preparation programs. An overview of the state of the art in this study courses has been considered as classical frameworks that

concentrated on quantitative as well as qualitative research. Finally, from this literature, a new data science foundation can be presented for education research. In (Chalaris *et al.,* 2014) the abilities of DM are based on the HE Institution (HEI) which manages to find novel explicit knowledge using DM models for EDM of Technological Educational Institute of Athens. Therefore, data applied in this approach is emerged from students' queries in classes by estimating the department of the educational institute. Unsupervised structure identification approaches like Clustering, Factor analysis, and Social network examinations have been utilized by EDM for identifying the new structures in diverse kinds of educational information. Along with that, association rule mining and corresponding types of pattern mining schemes are applied for identifying the relationships among various parameters in educational information. EDM is employed in different operations like developing an Intelligent Tutoring System (ITS) (Anderson *et al.,* 1985) in which distinct frameworks have been applied by ITS in order to establish student's recent knowledge by means of different skills. A massive number of technologies depends upon the knowledge tracking method and its difference. Followed by, many of the interesting operations such as automatic, data-based courses, college degree organizers as course suggestions, learning the effect of student's social hierarchy in academics when compares with many other applications.

EDM bridges among 2 modalities namely, Education and Computing sciences in which DM as well as ML models are considered to be the subfields of computing sciences. In EDM, DM and ML are assumed as subfields that are highly significant to show increments and advancements in academics and the teaching process. Therefore, uninformed by using educational strategy, CS, and details results in limited educational quality. The application of the system and CS in education is one of the trivial processes. Afterwards, computers are utilized as a tool for teaching in a drill-based method (Bates, 2015). In particular, it is considered as the height of behaviorism, and computers have resulted in effective outcomes. In addition, programmed learning structural data is capable to offer an immediate response for learners, and tests learning.

The application of machine depends upon the behaviorist nature termed as Computer Assisted Learning (CAL)/Computer Based Training (CBT). However, this mechanism was outdated, especially as it is not applicable to manage numerous learning processes like critical thinking, investigation, and synthesis which are essential in higher studies, even though CBT is applied in the educational cent. The exploitation of the system in education is "training", as it underscores a reputed approach where the first 3 phases of Bloom's classification of cognitive areas targeted such as knowledge, comprehension, and application. Hence, CBT is applied for accomplishing better efficiency when the perfect training is performed. The term "computer-mediated communication" was deployed in the last decades. Then, researchers have experimented with using computers in education which is categorized into 2 classes, with a focus on using computers "for programmed learning" or using the computers in order to interact with one another (ibid). Followed by, isolated computers have been employed. Then, Collection, Generation, and examining data is an arduous operation. Also, a revolution in the World Wide Web (WWW) has been implanted. The initial LMSs have been introduced in recent times (WEB-based Course development Tools (WebCT)). Online teachings as well as e-learning platforms are considered real-time environments. The time-consuming frameworks to load and explore the materials are limited significantly. Followed by, a complicated application of a system for learning the transcended in "training." The norm EDM has been employed in the past decades (Romero *et al.,* 2007).

A well-known approach for examining the supremacy of universities depends upon the excellence of students' academic performance. One of the popular and effective applications is the prediction in EDM which forecasts the students' GPA and educational performance, which is classified as excellent, very good, good, moderate, and so on. This kind of prediction is applicable in various universities which help in finding the best students for offering scholarships. Followed by, the literature has addressed the effect of successive features on predicting students' academic efficiency or GPA at the UG level: (Sembiring *et al.,* 2011) considered 300 students for predicting the final grade of students from faculty of computer systems as well as software engineering. The importance of a feature has been sampled by applying multi-variant analysis models. It is recommended that family support is highly important in predicting student's performance. Besides, students' interest does not have any impact.

III. Methodology

ACO-LR based Feature Selection and Classification Model

This section has developed an ACO based FS (ACO-FS) with LR based classification model called ACO-LR for the classification of EDM data. The proposed model involves a set of two major phases namely FS and classification (Figure 3).

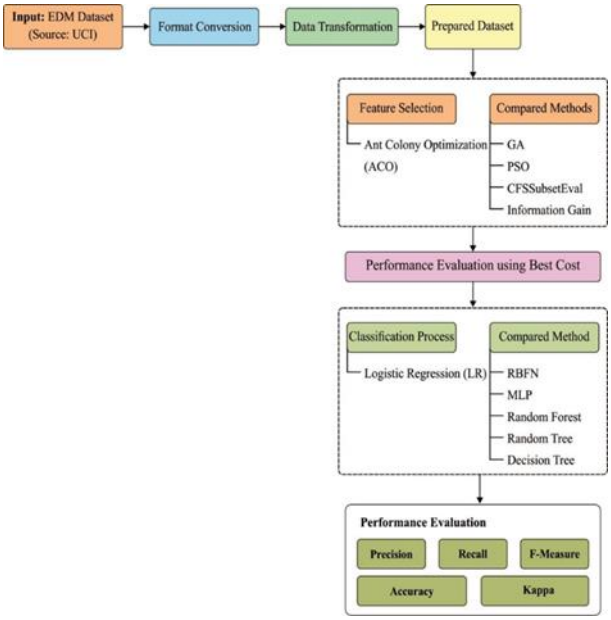

Figure 3. Block diagram of ACO-LR model

Preprocessing

In the ACO-LR model, preprocessing takes place in two stages namely format conversion and data transformation. In the beginning, the format conversion process takes place where the data in any other formats such as .xls will be converted into .csv files. Then, the data transformation process will begin where the data present in diverse formats in the dataset are transformed in an appropriately.

ACO Based Feature Selection

Here, ACO-FS is executed to pick up the feature subset from the educational data. ACO is a population-based metaheuristic algorithm which has the capability of searching the population in parallel. It offers a faster exploration of the optimal solution and adapts to modifications like new distance. Besides, the ACO algorithm has offered a better convergence rate. So, it is applied for the selection of features in the applied educational data. The intention of FS for ACO is to recognize the minimum feature count and to attain maximum classifier accuracy with the reduced processing expense.

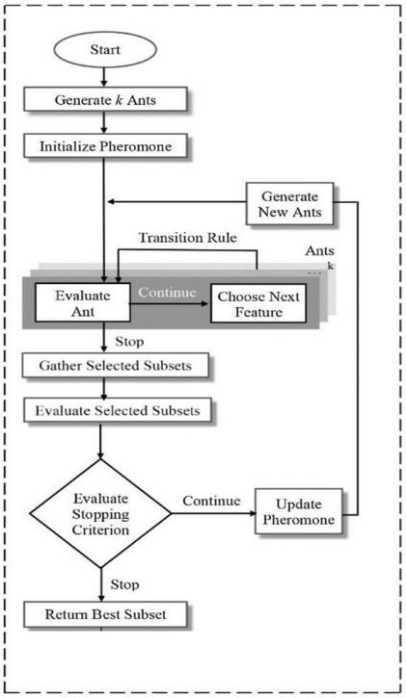

Figure 4. Flowchart of ACO-FS technique

LR Based Classification

Finally, the LR classifier gets executed to classify the feature reduced data to identify the class label in an appropriate way. The classification models mainly used for developing an approach for mapping the data to a specific class by means of existing data. It is employed for extracting required data items from this method for detecting a data movement. The dependent attribute of the LR scheme is a binary-classification. It means that the LR model is prominently utilized for solving 2-class problems. It is engaged in the LR based classification model (Figure 5). The major theme of this method is to predict whether a student is a fast learner or a slow learner, named as a binary- classification issue. Besides, the LR approach is applied in DM, automatic disease diagnosing, and economical forecasting while predicting health issues. At last, it has desired to uses LR as a projected model.

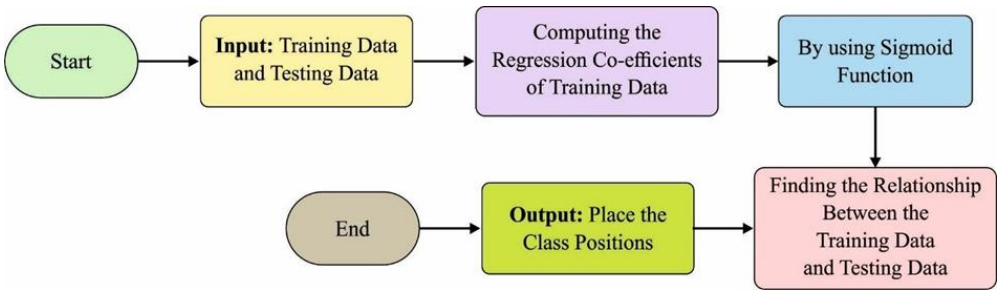

Figure 5. Process involved in LR based classification

The classification issues are the same as linear regression. The linear regression problem is applicable to forecast frequent value. The detected value of the classifier maybe 0 or 1, so-called as a critical point. If the result is 1, then the threshold value is higher, otherwise, it is 0. The final variable of LR should be between 0 and 1. LR is defined as a regression model that reduced the predictive radius and prediction value to [0, 1]. Moreover, it is added with a layer of the sigmoid function. The features are enclosed linearly and forecasted using a sigmoid function.

IV. Results And Discussion

Performance Evaluation of ACO-LR Model

Dataset Used: The outcome of the projected ACO-LR model is assessed on a benchmark dataset from UCI repositories. The dataset comprises a collection of 649 instances under the occurrence of 33 features. In addition, a set of 2 classes were present in the dataset, including a total of 549 samples under 'pass' category and 100 samples under 'fail' category (Table 4.4).

Table 1. Details of the dataset

Description	Values
No. of Instances	647
No. of Features	29
No. of Class	2
No. of Pass Samples	547
No. of Fail Samples	100

The distribution of the class variables along with the grade and their equivalent ranges (Table 4.5). The level I exist in the interval of 16-20 under the grade 'Very Good'. The level II present in the interval of 14-15 under the grade 'Good'. The level III is present in the interval of 12-13 under the grade 'Satisfactory'. The level IV is present in the interval of 10-11 under the grade 'Sufficient'. The level V exists in the interval of 0-9 under the grade 'Fail'. Then, the list of the features presents in the database and the attributes description.

Table 2. Class Variable Distribution (0-20)

Level	Grade	Ranges
I	Very Good	16 to 20
II	Good	14 to 15
III	Satisfactory	12 to 13
IV	Sufficient	10 to 11
V	Fail	0 to 9

Results Analysis of ACO-LR Based Feature Selection Process

The outcome of the FS process attained by the ACO algorithm (Figure 4.7). The outcome clearly shows the best cost attained under a set of 20 iterations. It is shown that the ACO-FS model has attained the best

cost of 0.030509. This value depicted the effective FS results of the applied model on the test dataset applied.

The features which are chosen by the different FS models on the applied dataset (Table 6). It is shown that the ACO-FS model has chosen a collection of 24 features from a total of 33 features with the best cost of 0.030509. At the same time, it is noted that the CFSSubsetEval and Information Gain models have shown poor results by attaining the best cost of 0.366000 and 0.386920 respectively (Table 4.6).

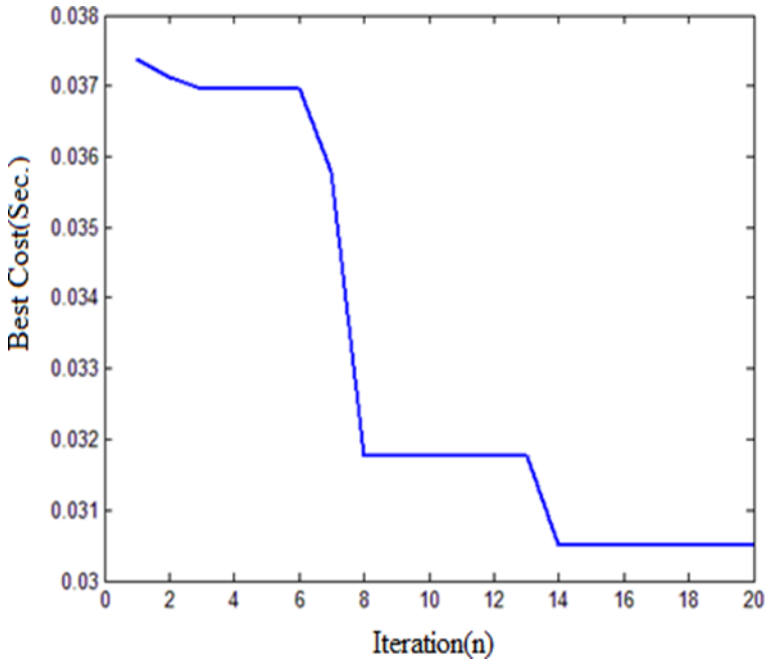

Figure 6. Feature selection results in ACO-FS model

Results Analysis of ACO-LR Based Classification Process

The confusion matrix generated by LR and ACO-LR on the applied dataset (Figure 7 (a) and Figure 7 (b)). It can be noticed that the LR model properly classifies a set of 536 instances as 'pass' and 79 instances as 'fail' category. But, enhanced results are offered by the ACO-LR model which properly classifies a set of 538 instances as 'pass' and 78 instances as 'fail' category.

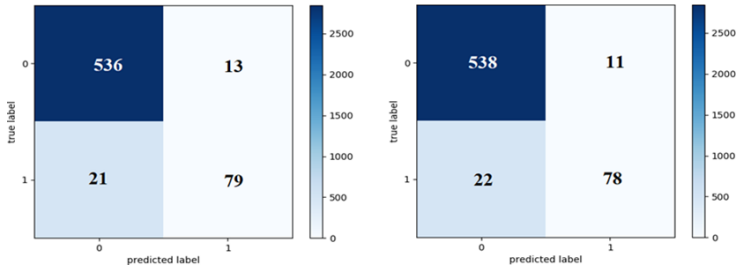

Figure 7. Confusion Matrix of LR and ACO-LR

The generated confusion matrices of the proposed and compared models. By the use of the values present in the confusion matrix, the corresponding classification results will be determined. The result analysis offered by the presented model on the applied dataset (Table 4.5). It can be observed that the ACO-LR algorithm achieved a higher precision, recall, accuracy, F-measure, MCC, and kappa value of 92.74%, 98.07%, 92.29%, 95.31%, 74.61%, and 73.67% respectively

Table 3. Confusion matrix of Proposed ACO-LR with Existing Methods

Methods	TP	TN	FP	FN
ACO-LR	538	78	11	22
LR	536	79	13	21
RBFN	510	71	39	29
MLP	519	63	30	37
RF	537	64	12	36
RT	505	64	44	36
DT	525	68	24	32

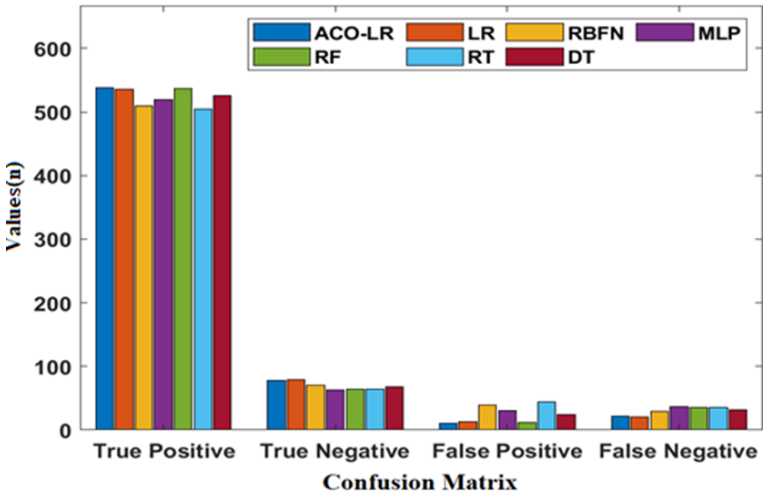

Figure 8. Confusion Matrix analysis of different models

In short, the simulation outcome portrayed that the ACO-LR technique has resulted in a superior precision, recall, F-measure, and kappa 97.99%, 96.07%, 97.02%, and 79.57% correspondingly. The above-mentioned tables and figures portrayed the superior characteristics of the presented ACO-LR model over the compared models.

V. Conclusion

The research successfully demonstrates the application of machine learning models, particularly the ACO-LR hybrid approach, for improving student performance prediction. By integrating feature selection with classification, the ACO-LR model achieved superior results compared to traditional methods, offering higher accuracy and efficiency in processing educational data. The study emphasizes the critical role of data mining and machine learning in addressing challenges such as early dropout detection and individualized learning pathways. The findings advocate for the broader adoption of predictive analytics in educational institutions to enhance teaching methodologies, improve student outcomes, and support evidence-based policy-making. Future research should expand on these methodologies to include diverse datasets and refine algorithms for greater scalability and adaptability in global educational contexts.

Reference

[1] Baradwaj, B.K. And Pal, S. (2012). Mining Educational Data To Analyze Students' Performance. International Journal Of Advanced Computer Science And Applications.,2(6):63-69.
[2] Bates, A.W. (2015). Teaching In A Digital Age: Guidelines For Designing Teaching And Learning. Tony Bates Associates, Isbn: 978-0-9952692-1-7
[3] Bharara, S., Sabitha, S. And Bansal, A. (2018). Application Of Learning Analytics Using Clustering Data Mining For Students' Disposition Analysis. Education And Information Technologies., 23(2): 957-984.
[4] Bogarín, A., Romero, C., Cerezo, R. And Sánchez-Santillán, M. Clustering For Improving Educational Process Mining. In Proceedings Of The Fourth International Conference On Learning Analytics And Knowledge., Pp. 11-15,2014.
[5] Bouchet, F., Harley, J.M., Trevors, G.J. And Azevedo, R. (2013). Clustering And Profiling Students According To Their Interactions With An Intelligent Tutoring System Fostering Self-Regulated Learning. Journal Of Educational Data Mining., 5(1):104-146.
[6] Bravo, J. And Ortigosa, A. (2009). Detecting Symptoms Of Low Performance Using Production Rules. International Working Group On Educational Data Mining.,4(2):31-40.
[7] Breiman, L., Freidman, J., Olshen, R. And Stone, C. (1984). Classification And Decision Trees. Taylor And Francis Group, Fl, Isbn:978-0-412-04841-8.
[8] Bucos, M. And Drăgulescu, B. (2018). Predicting Student Success Using Data Generated In Traditional Educational Environments. Tem Journal., 7(3):617.
[9] Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C. And Tsolakidis, Improving Quality Of Educational Processes Providing New Knowledge Using Data Mining Techniques. Procedia-Social And Behavioral Sciences, 147, Pp.390- 397,2014.
[10] Chandamona, P. And Ponperisasmy, D. (2016). Improved Analysis Of Data Mining Techniques On Medical Data. International Journal Of Nano Corrosion Science And Engineering., 3(3):85-90.
[11] Chrysafiadi, K. And Virvou, M. (2013). Student Modeling Approaches: A Literature Review For The Last Decade. Expert Systems With Applications., 40(11): 4715-4729.
[12] Cocea, M. And Weibelzahl, S. Can Log Files Analysis Estimate Learners' Level Of Motivation? International Conference On Learning Analytics, Pp.32-35,2006.
[13] Czibula, G., Mihai, A. And Crivei, L.M. S Prar: A Novel Relational Association Rule Mining Classification Model Applied For Academic Performance Prediction. International Conference On Knowledge-Based And Intelligent Information & Engineering Systems, Pp.20-29,2019.
[14] Daniel, B. (2015). Big Data And Analytics In Higher Education: Opportunities And Challenges. British Journal Of Educational Technology., 46(5):904-920.

[15] Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F. And Alowibdi, J.S. Predicting Student Performance Using Advanced Learning Analytics. Proceedings Of The 26th International Conference On World Wide Web Companion, Pp. 415-421,2017.
[16] Dekker, G.W., Pechenizkiy, M. And Vleeshouwers, J.M. (2009). Predicting Students Drop Out: A Case Study. International Working Group On Educational Data Mining.,*4*(10):41-50.
[17] Del Campo, C., Urquía-Grande, E., Pascual-Ezama, D., Akpinar, M. And Rivero, C. (2020). Solving The Mystery About The Factors Conditioning Higher Education Students' Assessment: Finland Versus Spain. Journal Of Education And Training.,62(6):617-630.
[18] Delavari, N., Phon-Amnuaisuk, S. And Beikzadeh, M.R. (2008). Data Mining Application In Higher Learning Institutions. International Journal Of Informatics In Education.,7(1):31-54.
[19] Demetriou, C. And Schmitz-Sciborski, "A. Integration, Motivation, Strengths And Optimism: Retention Theories Past, Present And Future," Proc. 7th National Symposium On Student Retention,201, Pp.300-312,2011.
[20] Dutt, A., Ismail, M.A. And Herawan, T. (2017). A Systematic Review On Educational Data Mining. Ieee Access., 5:15991-16005.
[21] Feo, T.A. And Resende, M.G. (1995). Greedy Randomized Adaptive Search Procedures. Journal Of Global Optimization, 6(2):109-133.